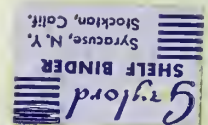


COMPUTER RECOGNITION OF SPEECH
UTILIZING ZERO-CROSSING INFORMATION

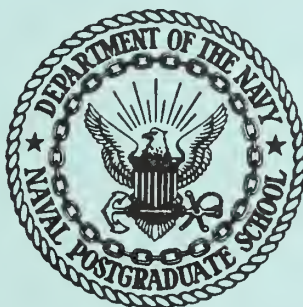
by

John Francis Taylor

UATE SCHOOL
F. 93940



UNITED STATES NAVAL POSTGRADUATE SCHOOL



THESIS

COMPUTER RECOGNITION OF SPEECH UTILIZING
ZERO-CROSSING INFORMATION

by

John Francis Taylor

June 1968

This document is subject to special export controls and each transmittal to foreign government or foreign nationals may be made only with prior approval of the U. S. Naval Postgraduate School.

LIBRARY
NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIF 93940

COMPUTER RECOGNITION OF SPEECH UTILIZING ZERO-CROSSING INFORMATION

by

John Francis Taylor
Lieutenant, United States Navy
B.S., Providence College, 1961

Submitted in partial fulfillment of the
requirements for the degree of

ELECTRICAL ENGINEER

from the

NAVAL POSTGRADUATE SCHOOL
June 1968

ABSTRACT

The nature of speech sounds is studied with particular emphasis on the information bearing elements of speech. The association of the amplitude clipped-speech zero-crossing rate, formant frequencies and information content of a speech signal is presented and capitalized upon to produce readily extractable first and second formants from the speech wave.

Various methods of processing the formants to generate unique patterns for particular sounds are attempted, with a time plot of the arithmetic difference of the two formants being explored in detail. The object being to obtain machine recognition of speech.

Control Data Corporation 160 computer machine language programs are prepared to realize an Euclidean comparison of spoken numbers zero to nine against a previously stored "dictionary." Testing showed this type processing satisfactory for some voices, but not readily extendible to many voices with the same "dictionary." Methods of overcoming this shortcoming are suggested.

TABLE OF CONTENTS

Section	Page
1. Introduction	7
2. Nature of Speech Sounds	7
3. The Spectrograph	13
4. Nature of Clipped Speech	14
5. Syllables versus Phonemes for Speech Recognition Schemes	18
6. The Vector Display	21
7. Formant one versus Formant two versus Time	25
8. Formant two minus Formant one versus Time	26
9. CDC 160 Computer and Algorithms	30
10. Conclusions and Recommendations	34
APPENDIX	
I. Block Diagram of Vector Display	42
II. Circuits used to extract Formants one and two from the Speech Signal	44
III. Patterns of Formant one versus Formant two versus Time	47
IV. Patterns of Formant two minus Formant one versus Time	50
V. Computer Programs	

LIST OF ILLUSTRATIONS

Figure		Page
1.	Typical Frequency Spectrum of a Laryngeal Tone Prior to Articulation	9
2.	Typical Resonance Pattern Produced by the Articulation Organs	9
3.	Spectrum of Laryngeal Tone after Articulation	10
4.	Effects of Infinite Clipping on a Speech Wave	16
5.	Block Diagram of Circuits used to Generate Analog of Zero-crossing Rate of Original and Differentiated Speech Signal	24
6.	Circuit for Generating Formant one versus Formant two versus Time	27
7.	Axes Directions for Plot of Formant one versus Formant two versus Time	27
8.	Block Diagram of CDC-160 Computer and Peripherals as used in this Work	31
9.	Block Diagram of Circuits used to Produce the Vector Display	43
10.	Circuit for Generation of Second Formant Frequency	45
11.	Monostable Multivibrator and Low-pass Filter as used on both Formant Channels	46
12.	Photographs of Formant one versus Formant two versus Time for numbers 2 through 5	48
13.	Photographs of Formant one versus Formant two versus Time for numbers 6 through 9	49
14. - 16.	Photographs of Formant two minus Formant one versus Time for numbers 1 through 9 by three male speakers	51-53

Acknowledgements.

The author extends his thanks to Dr. Gerald D. Ewing of the Electrical Engineering Department for his assistance, and especially for his willingness in allowing me to pursue this topic as I wanted. The enthusiasm and excitement generated by this personally guided journey into the unexplored was most enjoyable.

Professor D. B. Hoisington's help as second reader provided many improvements to the final draft. His assistance is sincerely appreciated.

I also thank Mr. Walter Landaker of the Digital Control Laboratory for his guidance in operation and programing of the digital computer.

1. Introduction.

In the present age of scientific discovery, man has become more and more dependent on the use of electronic computers. The future holds in store even more use of these devices with no apparent limit in sight. As this powerful tool becomes more universally important in man's day to day existence, it becomes increasingly more aggravating that he has to speak to it in its mode of communication, paper tape or punch cards; and not in his own, the spoken word. Even today, at the very infancy of the computer age, the time required to do many computations is less than the time required to instruct the machine in how to do them. This interface problem between man and machine promises to become even more severe as computers become more sophisticated.

All this points to the need of a method of achieving machine recognition of speech. Much work is now being done on this problem, but it is far from solved. The work described in this thesis is concerned with an approach to a simplified form of this problem, which may be a stepping stone on the path to its eventual solution.

2. Nature of Speech Sounds.

In order to obtain insight into the information carrying aspects of speech, it is well to study the nature of the speech producing process. Speech sounds are produced by modulations forced on the air stream coming out from the lungs. These modulations can occur first in the larynx, the first valve the air stream meets in its travel. The larynx is made up of bundles of muscle fibers, called vocal cords, which can be brought together to restrict the flow of air or to stop it completely. To produce a speech sound these folds are brought together to stop the air flow. When sufficient pressure is built up behind the

closed orifice to push the cords apart, a puff of air escapes. The cords then close until pressure again forces another puff out. This process, which occurs at the rate of a few hundred times a second, is called phonation. The nature of this sound production indicates that it is very much different from a pure sinusoid, perhaps more like a triangular wave in shape, showing that harmonics are present extending to frequencies much higher than the basic rate of phonation. By controlling the tension on the vocal cords the fundamental frequency of phonation can be controlled.

Following the larynx, the air flow, which can be referred to as the speech wave, passes into the vocal tract where the major part of the intelligence to be transferred by the speech process will be added. The vocal tract consists of the throat, mouth, and nasal cavity, and the process by which these cavities, joined with the lips, tongue and teeth produce the desired modulation is called articulation. To understand the effects of articulation on the speech wave it is first necessary to define the various types of speech sounds, since the effects of articulation, although similar in nature, are different in principle of information processing: sometimes adding information to a sound wave, sometimes producing the sound wave itself.

Speech sounds may be divided into two classes according to their origin of production: voiced sounds if they are produced in the larynx as described above, and later modified by the articulation process, or unvoiced sounds if they are produced solely in the organs which follow the larynx. If the frequency spectrum of a voiced sound were plotted as it appears out of the larynx and before any articulation has occurred, it would look something like figure 1. : a fundamental frequency



Figure 1. Typical frequency spectrum of a laryngeal tone prior to articulation.

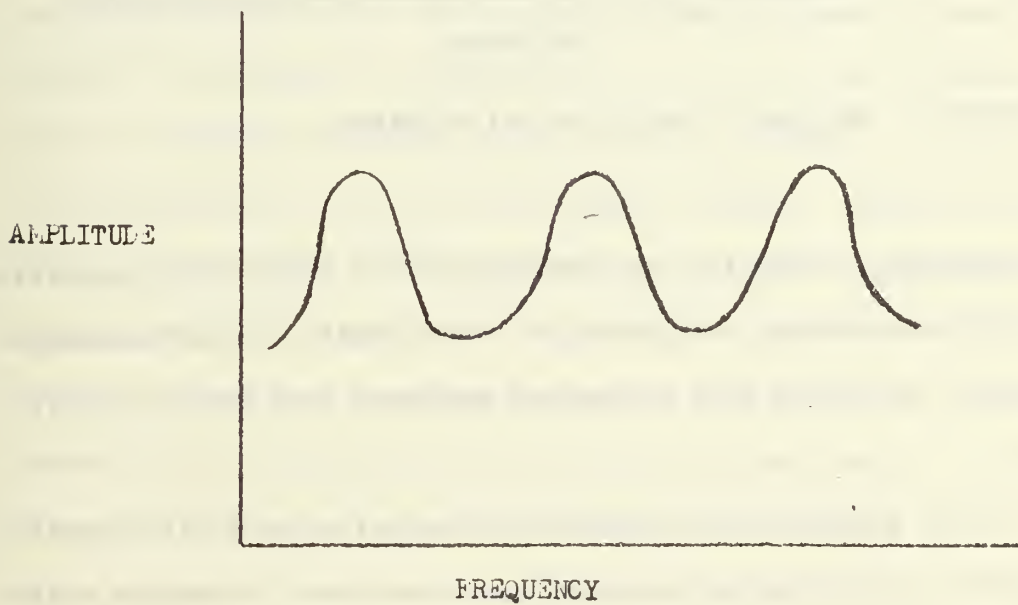


Figure 2. Typical resonance pattern produced by the articulation organs.

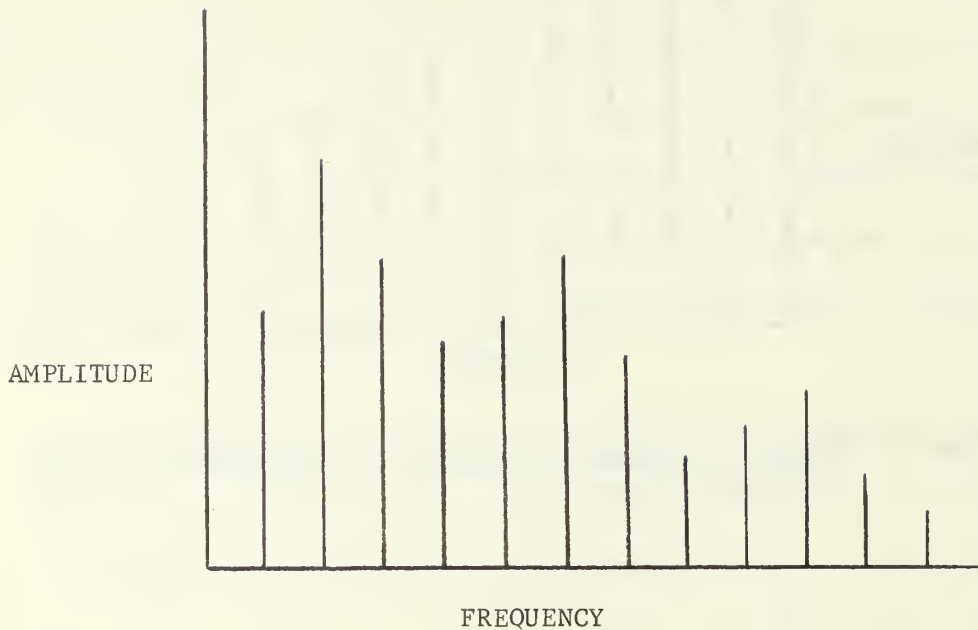


Figure 3. Laryngeal tone after articulation.

(corresponding to the rate of vibration of the vocal cords) occurring with large amplitude, and harmonics, or multiples of the fundamental frequency, occurring with decreasing amplitude with increasing frequency.

In the process of articulation on the voiced wave, the organs of the vocal tract are moved into different positions to produce various frequencies of resonance. These resonant frequencies serve to enhance the amplitudes of those frequency components of the voiced sound which fall in their regions. A plot of these frequency resonances might look something like shown in Figure 2.

The combined result of the articulation of Figure 2 on the voiced wave of Figure 1. would be a spectrum of amplitudes occurring

predominantly at those frequencies where the resonance humps occurred, with the amplitude of each hump being less than the preceding one as frequency increases (Figure 3). The selection of three resonance humps for this example was no accident, since this is precisely the way that articulation affects the voiced sounds in most instances. These three frequency regions where most of the amplitude, and hence energy, is located are called formants, and their location in the frequency spectrum is felt to have much to do with the information bearing mechanism of speech. Much more will be said of formants in the pages that follow.

Speech sounds which occur with no vibration of the vocal cords are called unvoiced sounds. It is apparent that a frequency spectrum model of such a sound would not be as simple as for the voiced sounds, and this is a problem that makes speech analysis difficult. No one model can be extended to all the various speech sounds. Indeed, we have only broken speech sounds into two major classifications and already are unable to describe them with a single model. More sub-classifications are yet to come, and they will be equally elusive when it comes to a common basis of modeling. It is to be noted that vowels, all of which are voiced and steady state type sounds in nature, fall into one class, while consonants, which are more transient in nature, occur in both voiced and unvoiced classes.

Consonants are commonly classified according to the way they are produced. They are generally divided into six categories as follows:

Plosives or Stops - These consonants are produced by a stopping and then sudden release of the air. The stop plosives are p, b, k, and g.

Continuents - Continuents, unlike plosives, may be continued or prolonged during a breath. They are further sub-categorized as nasals, laterals, and fricatives. Nasals are produced by stopping the air in the mouth and releasing it through the nostrils. Laterals are produced by placing the tip of the tongue on the upper gum ridge and releasing the air over the sides of the tongue. The only lateral in the English language is l. Fricatives are formed by forcing the air through a very narrow opening in the articulation organs. The fricatives in English are f, v, th, r, h, s, z, sh, and zh.

Glides - Glides are characterized by a continuous movement of an articulation organ as the sound is produced. The glides are w (we), wh (when), and the initial sound in yes.

Vowel like consonants - These consonants are so named because they have some of the characteristics of the vowels. They are w, r, l, m, n, ng, and y as in yes.

Glottal sounds - Glottal sounds are sounds produced in the glottis, the opening between the vocal cords. The only glottal sound in English is h.

Affricatives - Affricative sounds are plosives followed immediately by fricatives. The affricatives include ch and j. There are many other ways to classify speech sounds and also other groups within the classification here which have not been included. This breakdown is not meant to be exhaustive, but rather only complete enough to make the

reader aware of the definitions of these terms as used in this paper, and the degree of difference in speech sounds this researcher had assumed in undertaking the work described later.

3. The Spectrograph.

In trying to machine recognize speech, the problem is one of finding elements or parameters of the spoken word that are uniquely characteristic of that word and no other. Individual speaker characteristics can be regarded as noise and not of interest. To be sure, emphasis, timing and so on, can affect the meaning of what is being said, but at this stage in the development of speech recognizers, the simpler problem is sufficiently difficult to warrant study.

There has been much work done in the area of trying to extract the informational content of speech from the "noisy" form in which it appears from the speaker's mouth. In particular, the objective has been twofold: To reduce the bandwidth required to transmit the information, and to provide a visual presentation of the information. The former has resulted in various kinds of vocoders such as formant vocoders, correlation vocoders, fixed channel vocoders, and hybrid combinations of these (12). The latter work has been chiefly concerned with the sound spectrograph (21). The spectrograph is of particular interest to this work and so a brief discussion of it follows.

The sound spectrograph was first presented in the literature in "Science," November 1945. It is essentially a device for making paper strip recordings of frequency and intensity versus time for short sound samples. The recordings are so made that the variations of vocal resonances (formants) with time are displayed conspicuously. Spectrograph recordings of vowels (Ref. 21) show the formants as well defined

bars at specific frequencies. For the consonants the only way to determine where their formants are is to see where the lines came from in transitioning into the vowels.

By studying a few such spectrograph recordings, one can see that the locations of these formant lines is a characteristic of a particular sound. In particular, the second formant bar is considered to be of special importance. It would then be reasonable to seek to use this formant information for a speech recognition scheme except for the problem of identifying the consonant formant frequencies. The spectrograph has presented vividly the importance of the formants in finding the information carrying elements of speech, but just has not provided a means of getting a hand on all these formants easily.

4. Nature of Clipped Speech.

If a speech wave is viewed in the time domain, that is, a plot of amplitude versus time, an obvious characteristic is the great dynamic range of amplitudes that are present. Variations of up to 60 db are not uncommon in normal speech. In particular, it is noted that vowels are on the average 12 to 28 db higher than the consonants. This wide dynamic range of normal speech presents problems in speech processing for transmission, since a transmitting system would have to work at a very low average power (and, of course, lower range) if the exact shape of the speech wave were to be preserved. In order to increase the average power, work has been done in the area of speech clipping. The approach to this problem was to see how much peak clipping could be accomplished without distorting the signal beyond comprehensibility.

It has been found by various researchers that clipping the original speech waveform up to 12 db has no noticeable effect on the quality

and intelligibility. Clipping of about 12 db sounds as if the speaker were enunciating carefully.

The improvement in intelligibility for 12 db of clipping is somewhat surprising at first since the speech wave has definitely been distorted considerably, and one would expect degradation in performance. Actually, by reducing the peaks which are primarily associated with the vowels, the process serves to enhance the relative power in the consonants. Since the consonants are much more transitory in nature than the vowels, it is appealing to say from an information theory point of view that they are the primary information bearing elements in the signal, and to increase their relative power is to increase the emphasis on the information content of the speech wave (22). It also follows from this that the individual speaker characteristics are contained more in the vowels than in the consonants. One would, therefore, expect clipped speech to be somewhat less indicative of speaker voice traits, and this is an experimentally proven fact.

Pushing the concept of clipped speech to the absolute limit, a group at Harvard University studied the effects of infinitely clipped speech. Infinitely clipped speech being produced by clipping, amplifying, and reclipping until the only information contained in the processed wave is the places of time axis crossing, referred to as zero-crossings. An example of such an infinitely clipped waveform is shown in Figure 4.

From Figure 4 it can be seen that the amplitude information has been totally removed from the speech wave. It was found that despite this severe distortion, the clipped wave was 90% intelligible in the absence of noise. By differentiating the original speech wave prior

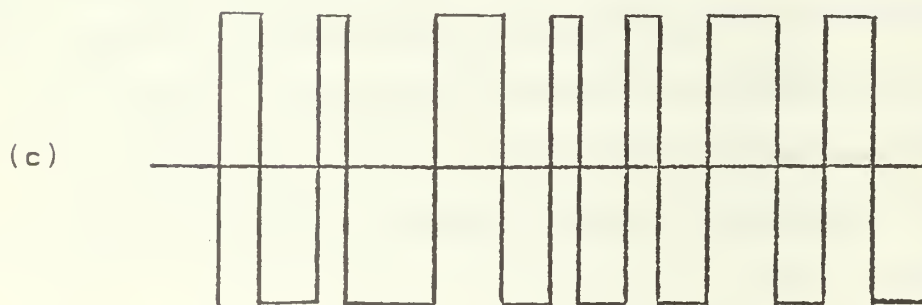
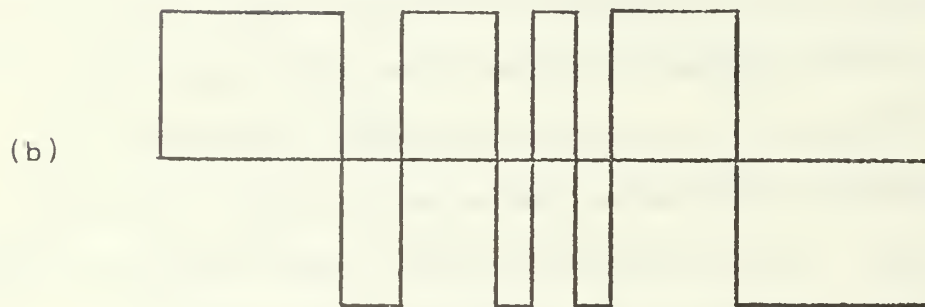
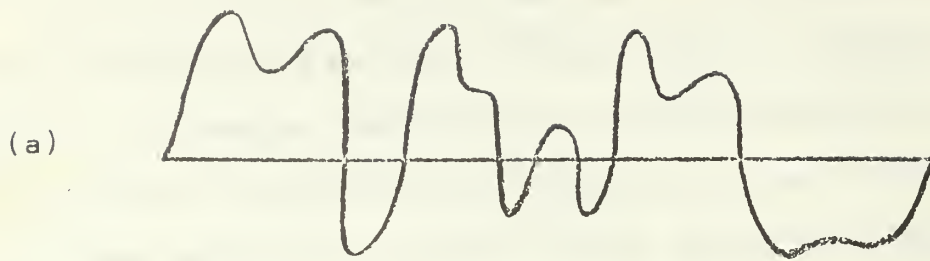


Figure 4. (a) Original waveform; (b) After infinite clipping; (c) Differentiated prior to infinite clipping.

to clipping, that is, producing an infinitely clipped wave whose zero-crossings correspond to the maxima and minima of the original wave, a 95% intelligibility was noted. This seems to indicate that a greater amount of information is contained in the higher frequency components, which are brought into the forefront when emphasized by the six db per octave increasing amplification of higher frequencies produced by the differentiating action.

The reason that clipped speech is still intelligible can be further seen from a frequency spectrum point of view. It is a basic fact of nature which is easily proven that the human ear is quite insensitive to phase. It has been common to assert that the information content of a speech wave is contained in the energy spectrum of its various frequency components. If the relative phases of these components are varied, within limits, thereby producing a wholly different amplitude versus time pattern, the ear would notice no difference.

The importance of clipped speech to the work undertaken in this thesis is the relationship of its zero-crossing rate to the formants of the speech sounds. Chang, Phil and Essigmann in their paper "Representations of Speech Sounds and some of Their Statistical Properties" (5) have demonstrated mathematically that the average rate of zero-crossing of the undifferentiated speech wave is very nearly a measure of the first formant frequency. Furthermore, the average rate of zero-crossing of the differentiated wave is a measure of the second formant frequency.

From the above interrelation of clipped speech and the first two formants, which are strongly believed to be the information bearing elements of the speech wave, it is hypothesized that equipment could

be designed to obtain the formant frequencies via the clipped speech zero-crossing rate. These formant frequencies could then be used together, or perhaps with the assistance of some other speech parameters to distinguish spoken words for at least a limited vocabulary and for a variety of speakers. This hypothesis is based on the appealing assumptions that the formant frequencies do contain the information and that the individual speaker characteristics can be eliminated by going to the formants via the clipped speech zero-crossing approach. This researcher has found no indications in the literature, save that discussed below (22), to indicate that anyone else has attempted to verify Chang's mathematical conclusions for speech sounds. The work below pointed out some distinct possibilities, and it is from there that the work for this thesis began.

5. Syllables Versus Phonemes for Speech Recognition Schemes.

In the initial phases of designing a scheme for the automatic recognition of speech, the question of how large a speech segment is to be analyzed at a time has to be considered. It has been suggested by several researchers and institutions engaged in work on this problem that the logical approach is to go to phoneme recognition, a phoneme being the smallest element of a speech sound. Since there are only 40 phonemes in the English language, this would lead to a minimal stored dictionary, and would be capable of responding to any word, even those that are not in existence at the time of the design of the device. However, the problems associated with analyzing an utterance as small as a phoneme make such a seemingly optimal approach difficult to implement. To be sure, some successes have been achieved in this method, notably by the Radio Corporation of America in their work on speech recognition for the Air Force.

Their work was directed towards the regions of decreasing and increasing spectral energy, rather than the energy peaks (formants) themselves. These features were found to be more easily abstracted and more invariant for their processing method. The use of phonemes was found to be satisfactory in their study of segmented speech. It is pointed out, however, that for continuous speech some provisions will have to be made for the changes that occur in the sound of phonemes caused by the neighboring sounds.

An equally large group has espoused the syllabic approach to the recognition problem, a syllable in this sense not necessarily meaning the same thing as a syllable in grammar. Estimates of the number of different syllables needed in a dictionary to adequately cover the English language run from 1000 to 2000, with those who embrace the phoneme approach voicing the latter figure. The actually needed number probably lies somewhere between, but it would seem that something considerably more limited could be used for most applications if and when a method is perfected. The phonetic typewriter developed by RCA Laboratories (18) is an example of a working model using syllabic recognition successfully for a vocabulary of 100 syllables. This system operates on an input of syllables or monosyllabic words spoken one at a time. These utterances are then normalized and their frequency spectra extracted by banks of filters for comparison with previously stored "dictionary" spectra. The authors make the point that the syllabic approach was chosen for their work because the sounds of the various phonemes have different characteristics when taken out of context, and are thus not felt to be a reliable indicator of the information in themselves, but only as they exist in the syllables.

To be sure, this syllabic approach is also a simplification of the overall problem since they themselves are known to be affected by the sounds that precede and follow. Any speech plan that does not set its sights on the problem of recognizing sounds as they occur in connected speech is never going to be a completely satisfactory all word recognizer. To solve the problem in any other form is to deal with it out of its natural environment, and hardly extrapolative into the more general case. However, the problem is complex enough at this stage of its study to warrant much more work on special situation type considerations until more is learned of the information carrying modes of speech. The investigation conducted by this student has encompassed just such a limited approach to the wider problem by restricting the study to monosyllabic words with a few minor exceptions.

The general aim of this work was to investigate experimentally the formant zero-crossing association discussed by Chang, et al (5) and to explore the possibility of using these parameters alone or with others to achieve patterns or matrices independent of individual speaker characteristics and highly indicative of the word being spoken. In the event that the parameters obtained were not suitable for this objective, it was proposed that the methods planned be used for analyzing individual utterances such as fricatives, plosives, and so on to determine if there is any correlation between data for just some such particular sounds. It is possible that the information contained in the formants, or, if you will, in the zero-crossing rate is only derived from particular articulations and not from all. The information obtained in the literature by this researcher indicates that there are no real definitive answers available in this regard and it is

unknown as to just what are the most important information bearing elements in the speech communications system. The systems of speech analysis and synthesis now existing have gained what successes that they have not so much from an application of scientifically applied knowledge, as from an engineering trade-off of bandwidth for a conglomeration of other characteristics of the speech wave, which somehow, through the benevolence of a sympathetic diety, has worked. The work described herein is to be considered as just one more such flail at this elusive problem.

6. The Vector Display.

In a report on a government sponsored research effort on signal processing by infinite clipping conducted at Georgia Institute of Technology in late 1963 and early 1964, B.O. Pyron and F.R. Williamson, Jr. discussed a vector display unit which they had developed for visually displaying voice and other short time, highly transient signals. They found that an analog signal proportional to the short-time running average of the zero-crossing rate of the original or differentiated speech wave was quite similar for the same sound by many speakers and distinctly different for other sounds. Another analog signal was produced proportional to the smooth envelope of the amplitude of the original waveform and used as a second coordinate for an oscilloscope display. This display, called a vector display by the originators, consisted of the averaged zero-crossing analog applied to the vertical deflection plates and the amplitude analog applied to the horizontal plates of a storage type oscilloscope.

The authors reported the patterns produced by the vector display had a tendency to correlate well for spoken words and seemed

independent of individual voice characteristics. In particular, the patterns produced using the differentiated waveform seemed to give the most distinctive shapes. This is plausible since the differentiated waveform is felt to carry more intelligence than the original for infinitely clipped speech.

As a beginning point in this thesis, the circuitry discussed in Pyron and Williamson's paper was constructed and their vector display studied. The circuitry used was exactly as presented in their report with the exception of minor corrections of obvious typographical errors. This circuitry is presented and discussed in Appendix I.

Patterns produced were similar to those in the reference. Utilizing a Hughs Memoscope to hold the highly transient characteristics of the analogs for study, patterns were generated for the numbers zero through nine by several male speakers. The objective being a series of distinctly different patterns for different numbers, but reasonably alike for the same number by various speakers. If such could be achieved, the ultimate objective being to use a digital computer for recognizing the patterns as the numbers they represent.

In working with the vector display it was noted (as reported by the originators) that the patterns were greatly affected by channel gain, bandwidth, and the time constant of the output low-pass filter. Therefore, in comparing patterns from day to day it became very important to insure that the precise same conditions existed for all the subjects in question. In particular, the level of amplitude of a speaker's voice at the microphone was most difficult to control, and this was noted to have an adverse effect on some patterns. However, such speaker voice power variations will have to be allowed for in any

practical system, and it is felt the amount of precaution taken here was sufficient to give the vector display a fair opportunity to prove itself up to the requirements of a machine recognizer of speech. Unfortunately, the results obtained in this study did not show this type pattern either unique enough for individual sounds nor consistent enough for the same sound by various speakers. To be sure, some sounds do have quite distinctive features, in particular those containing plosives or fricatives such as ship or tooth, but there seemed to be too many exceptions to make such a system workable.

A more recent paper on infinitely clipped speech by W.A. Ainsworth (1) pointed out that clipping systems which do not maintain a distinction between the polarity of zero-crossings provide less information than those which do, since to measure the frequency of zero-crossing in both directions is to measure the even harmonics of the wave only, thereby producing an harmonically distorted output. Intelligibility tests showed at least a 20% increase in intelligibility achieved by marking only the zero-crossings in one direction with pulses of the monostable multivibrator (See Appendix I).

With Ainsworth's results in mind and experience gained with the vector display, new circuitry was constructed to generate the analog of the zero-crossing rate of both the differentiated and original speech waveforms. A block diagram of this circuitry is shown in Figure 5 with the actual circuits outlined and their operation discussed in Appendix II.

Results obtained from this processing were somewhat more encouraging, but not enough to change the original opinion of this type display. Various input bandwidth and output low-pass filter time constant values

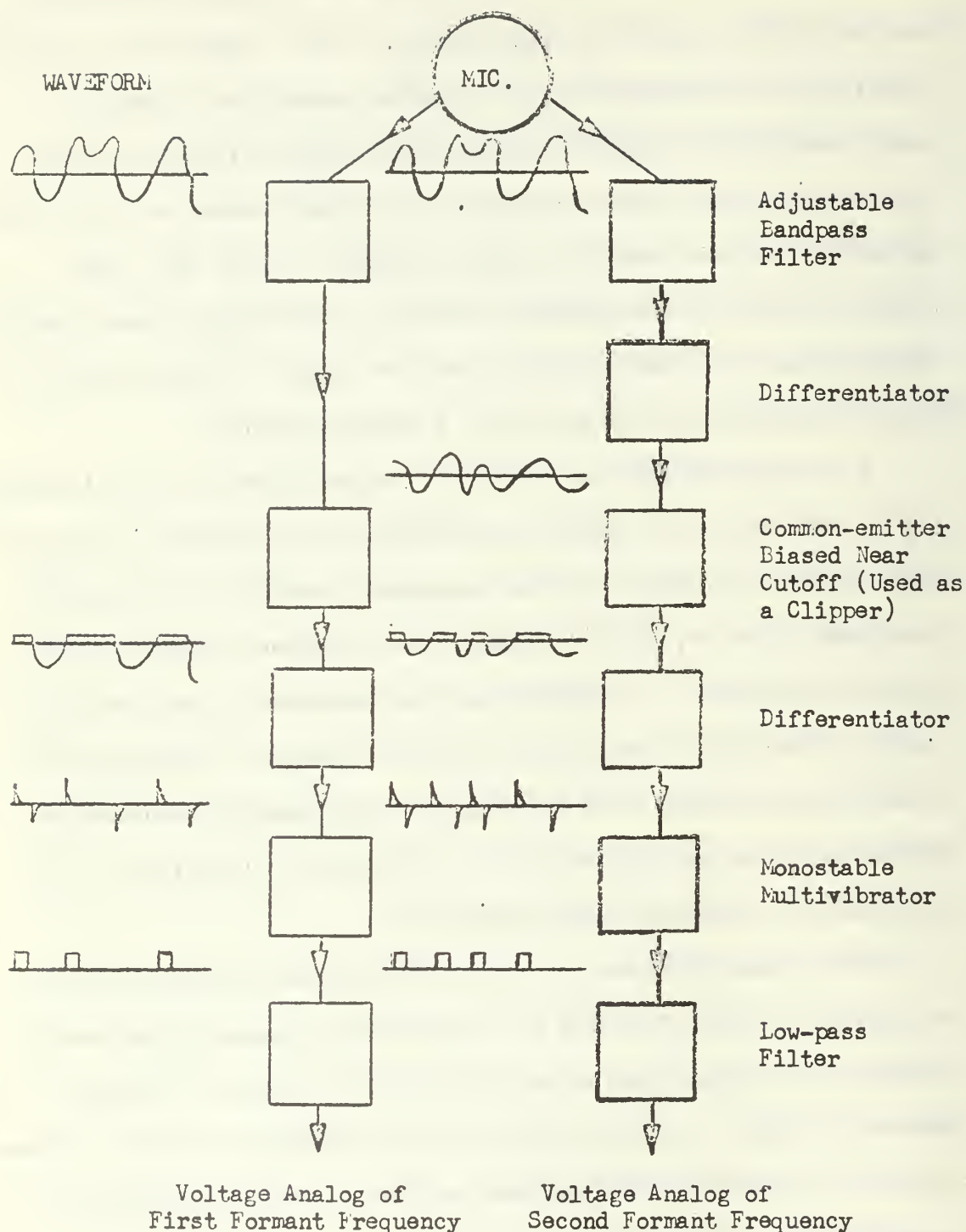


Figure 5. Block diagram of circuits used to generate analog of zero-crossing rate of original and differentiated speech signal.

were tried in the hope that the variations thereby produced in the patterns would be more severe in some than in all, and make them distinctive enough for further study. It was concluded on the basis of several days work in this approach that no real progress was being achieved and the vector display was abandoned.

7. Formant one versus Formant two versus Time.

In the course of working with the vector display the idea suggested itself that a display of averaged zero-crossing rate of the original waveform versus that of the differentiated wave should be of greater interest. This type display was appealing for the following reasons:

- 1) It would be a pattern defined by the first and second formants of the speech wave (as derived theoretically and demonstrated experimentally by Chang, et al in Ref. 5) which are held to be the information carrying elements of the speech wave.
- 2) It would eliminate the amplitude parameter from the somewhat promising vector display, a parameter whose phase dependence made its value suspect from the very beginning of this work.

Memoscope displays of first formant versus second formant were next generated for the numbers zero to nine by several male speakers. It became immediately obvious that this type pattern, although most promising in theory, left a trace that was too confusing for worthwhile analysis. It was apparent that if anything useful was to be obtained from this combination, another parameter would have to be included to spread the formant versus formant excursions of the trace out more from the origin of the axes.

Time was the obvious other parameter chosen and it was included in the present plot by applying a time sweep to both axes of the memoscope,

producing a time sweep diagonally rising across the scope. The circuitry used to achieve this is shown in Figure 6. The passage of the formant analogs through the amplifier stage in the circuit of Figure 6 caused 180° phase shifts and the resultant three dimensional plot is now as shown in Figure 7.

Patterns obtained for this type processing were very promising. With no bandlimiting on the input waveform, patterns were for a given speech sound very consistent for a variety of speakers. For the numbers zero to nine there was a need for more individuality in some of the patterns, especially those not containing plosives, fricatives, or stop consonants. Bandlimiting of the input wave to either or both channels offered possibilities of improvement, as did asymmetrical weighting of the formant channels. The second formant is felt by many speech researchers to be the principal carrier of information and so it seems reasonable to give it more emphasis in this kind of plot. More work was not done with this type display because the simpler approach to be discussed next gave much more interesting results at the same level of investigation. Typical pictures of the traces obtained with the display just mentioned are shown in Appendix III.

8. Formant two minus Formant one versus Time.

While working with the processing method just discussed, it became apparent to me that the display being studied was a vector sum of the two formants with time (not mutually orthogonal vectors). A simple arithmetic difference type process had been overlooked and with no justifiable reason. Such a combination would have the effect of canceling the similar portions of the formants, that is, the portions that are similar at the same time. Since the Hughes Memoscope utilized

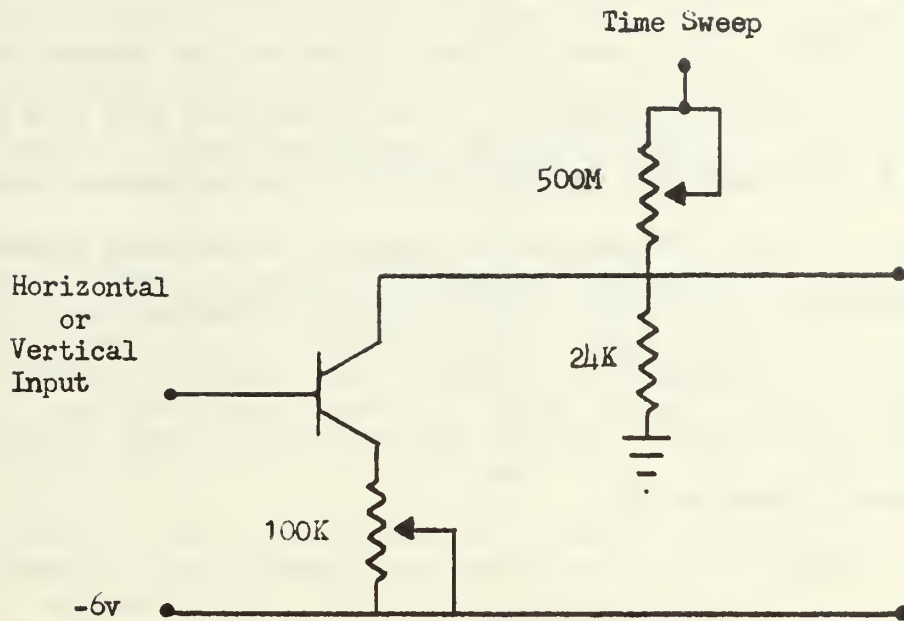


Figure 6. Circuit used for generating formant one versus formant two versus time plot.

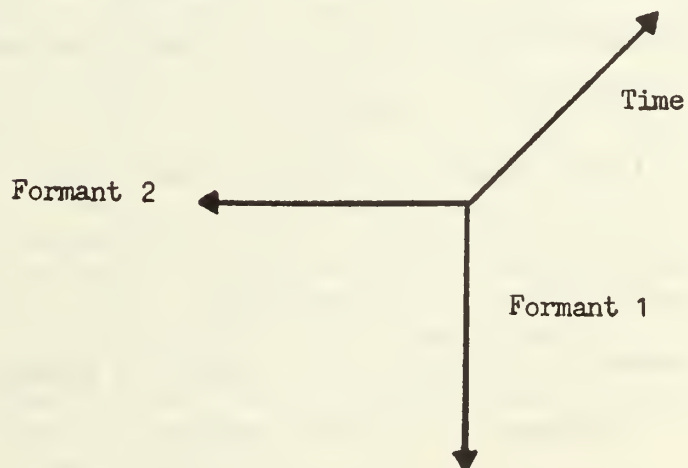


Figure 7. Axes directions for plot of formant one versus formant two versus time.

had a high gain differential preamplifier available as a plug-in unit, it was no problem to implement such a display.

Patterns were generated and studied for the numbers zero through nine by three male speakers. It was noted that this type pattern had to a fair degree the desired simplicity and uniqueness needed for the goal of machine recognition of speech. The patterns showed very good consistency for a given sound by various speakers, and, although not absolute uniqueness for different sounds, enough variance to make more work here feasible. Typical patterns for this kind of processing are shown in Appendix IV.

Rather than rely on visual consideration as done previously, it was decided at this stage of the investigation to feed the analogs being generated into a computer for comparison. Since such work would be best accomplished in a real time environment, both from the point of view of study and as an ultimate machine recognition capability, it was decided to do this work on a small, but more readily available computer, the Control Data 160. To be sure, forsaking the capabilities of the larger computers available here at the Naval Postgraduate School, the IBM 360 and the SDS 930, required greater effort in programming and provided a lesser degree of potential operations. However, for testing and evaluating a system such as this the advantage of working in a real time situation cannot be overestimated. Also, the value of any processing scheme is increased if the amount of computer capacity needed is held to a minimum, an inherent requirement here.

In selecting a method of pattern comparison for the computer to execute, it is immediately suggested to one who has studied some communications theory that a correlation technique is to be used.

Cross-correlation is defined as a graph of the similarity between two waveforms as a function of the time shift between them. However, if cross-correlation is considered as a matched filtering process, which it is, then the uselessness of this method in this work becomes apparent.

When a signal is cross-correlated with another, it is equivalent to an autocorrelation of that signal with itself plus noise. The effect is that the process acts as a filter and only allows through those frequencies which are in the signal. Thus, this method of signal processing is very powerful where you have a high frequency signal buried in wideband noise, such as the radar problem, but of little use when the signals of interest are bandlimited to below 300 Hz. As a comparator, correlation gives an average measure of the similarity between two waveforms. It is quite insensitive to local differences in the amplitudes of the two waveforms. Since local differences of the analog waves generated are the precise means by which I have attempted to perform machine recognition, correlation techniques would not work.

The poor performance of correlation in a low frequency problem has been experimentally shown by W. Bezdel in his paper regarding recognition of vowels by computer program using zero-crossing data (3). He noted poor results using correlation methods, although he does not explain why. A little thought on the matter makes one realize it would have been an anomaly if his results had been good, since the tool has little power at these frequencies.

Bezdel and Chandler indicated that they had success in their comparison work using a Euclidean distance measurement, that is, a point by point difference calculation between corresponding points on the

"unknown" vowel and a previously stored "dictionary" vowel. The dictionary word yielding the least total difference from the unknown would be selected as the best comparison. Such a method seemed of great interest to this researcher since it was simple in concept and therefore in keeping with my personal philosophy that "if whatever you are doing is complex and unwieldy, it is probably also wrong." This technique was also readily programable within the limitations of the CDC 160 computer.

The scheme employed in implementing this computer comparison, as well as a discussion of the CDC 160 computer and peripherals used is contained below.

9. CDC 160 Computer and Algorithms.

The Control Data Corporation 160 computer (see Figure 8) used is a parallel, single address electronic data processor controlled by an internally stored program in sequential locations. Memory capacity is 4096, 12 bit binary words. Instructions are executed in one to four storage cycles, with the time varying from 6.4 to 25.6 microseconds. Instructions could be either manually inputted via finger controls on the console face or by paper punch tape. Data can be inputted as above or from externally selected equipment.

The CDC 163 magnetic tape unit provides the capability of operating with many more than the 4080 memory cells contained in the computer by allowing you to dump information that is not immediately being used onto the tape and thereby freeing more memory locations for use. Tape stored data can be recalled at any later time.

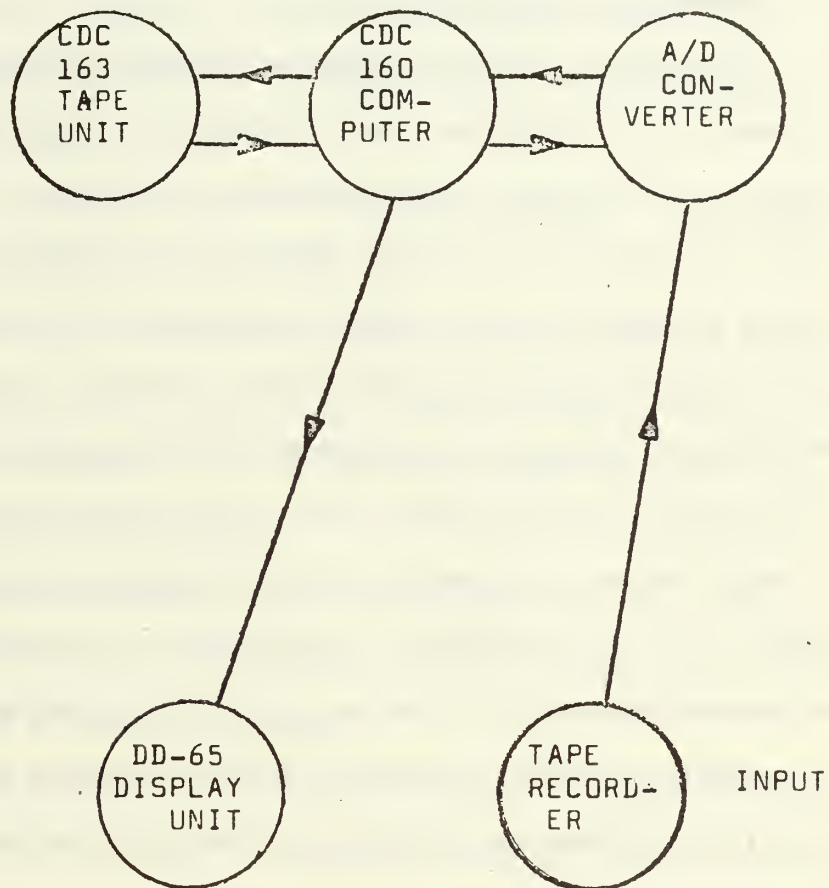


Figure 8. Block diagram of CDC-160 computer and peripherals as used in this work.

The DD-65 remote display unit provides a rapid means of trouble shooting programs in assembly language prior to utilizing a library assembler program to prepare the biocatal tape required by the CDC 160. This unit was also used to display the error versus time graphs during the comparisons as discussed later.

The analog to digital converter utilized is a non-commercially produced unit, having been constructed here at the Naval Postgraduate School by the Digital Control Laboratory Personnel. The A/D conversion unit is basically a multiplex sampling system which can sample one input at a time or up to 12 inputs in multiplex. By sampling each signal at the Nyquist rate (twice the highest frequency present) or higher, the digital samples will contain all the information that existed in the original analog waveform. The signals sampled here were limited to below 500 Hz so a sampling rate of 1 KHz was used in all my work. There was a time difference of approximately 100 microseconds between corresponding samples of the two inputs due to the multiplex nature of the sampling process. This error could be reduced somewhat by more judicious programming, but the error was not considered sufficiently serious to warrant changes at this time.

In order to make a Euclidean comparison, a program was written to perform the following operations:

- 1) Have the computer sit waiting for a 0.2 volt threshold before commencing requests for samples from the A/D converter. When a signal greater than the threshold was sensed, the computer would recognize that there was a word being inputted and would receive 512 samples each from both the first and second formant channels. The number 512 was chosen since it is sufficient to cover the time duration of the words

zero to nine used in this work, and also because it transforms to 1000 in the octal number system used by the computer. The DD-65 display unit has the capacity for only a 1000 point plot and so any larger amount of sampling would spoil the use of this analysis tool.

- 2) Take the difference of the second formant samples minus the first formant samples with the appropriate signs and store them on magnetic tape in the CDC 163.

- 3) Return the computer to thresholding operation until another 0.2 volt signal appears at the input.

After the numbers zero through nine have been stored on magnetic tape by the program described above, a second program is used to make the necessary comparisons to determine what unknown is being spoken. This program works as follows:

- 1) The computer waits for a 0.2 volt signal as above, and upon sensing it takes samples and subtracts the first formant from the second, storing the difference in memory.

- 2) The CDC 160 calls for the first stored word from the CDC 163 magnetic tape unit and takes the absolute value of the difference between that word and the unknown word, sample by sample, and stores this error sum in memory.

- 3) Each of the remaining nine words stored on tape are subsequently called into the computer for comparison and their individual error sums stored in memory.

- 4) When all ten words from the dictionary have been compared to the unknown word, the error sums are compared, and the word that has the lowest error sum is taken to be the best fit to the unknown, its corresponding number then being shown in the register of the CDC 160 as the number spoken.

If this monosyllabic recognition technique proved fruitful, one could extend it to polysyllabic words by some method of presorting according to word length or syllable count. Either type of classification could be accomplished without much additional computer memory requirements. The larger vocabulary could also be stored on magnetic tape where there is ample storage capacity.

To further assist in the analysis of the comparison work, the DD-65 display unit was programmed into the process to display a plot of the error versus time for each of the ten comparisons being made. This allows one to see on a real time display which words were close to the unknown and should not be, as well as which part of the correct word was wrong and caused the comparison not to be satisfactory. Such a display proved very beneficial in the subsequent work of trying to adjust the circuits used to improve performance.

The computer programs described above are contained in Appendix V.

10. Conclusions and Recommendations.

The numbers zero through nine were recorded by five male speakers. Computer recognitions were tried using each of the voices as a dictionary. Results were excellent for the same voice against itself, as would be expected. Attempts at inter-voice comparisons were not consistently successful for certain of the numbers as discussed below:

As can be seen from the pictures in Appendix IV, certain numbers are very unique in their shape, and as such are easy to match for many different voices. Examples of this are the 6, 7, and 8. It was possible to identify these numbers with any one of the five voices as the dictionary.

The remaining numbers are different enough to provide good identification if the speakers speak normally and clearly. It was noted that people frequently try to speak very clearly (and usually so much so that it is unnatural) when asked to speak into a microphone for testing. This presented a problem in identifying the number three. Some speakers said "th/-ree" and this gives a different pattern from the monosyllabic version of the word. Aside from such anomalies, results were very good, especially if each speaker heard how the others pronounced the words.

From the above testing it was apparent that some work would be required to make the patterns that were close in shape more unique, so more leeway might be allowed for individual speaker mannerisms. The input bandwidth was varied for each channel with the hope of increasing the differences between patterns. Since the first formant is expected to exist somewhere below 1 KHz, this channel was bandlimited between this frequency and 300 Hz. The second formant exists somewhere between 800 and 4000 Hz, and so this channel was limited to this frequency range. Other bands were tried also, but these settings seemed to do as well or better than any others and are reasonable for the parameters being extracted.

The output low-pass filters were also varied, with cutoff frequencies ranging from 300 Hz to 1000 Hz. Optimal settings for both channels seemed to be at 500 Hz.

Under these new conditions results obtained for the same five voices as above were improved, but problems still existed for the numbers 1, 4, 5, and 0. The error display indicated that the real key to discriminating between the patterns rests on the substantial excursions caused by the plosives, affricatives and fricatives, and those words

which contain none are inherently in trouble. With practice all the speakers tested began saying even these words the same and higher recognition scores were realized. In this regard the device functioned well as a speech training aid because all found it quite easy to enact the speaking enunciation of the best speaker and to obtain his good patterns. For those people who speak clearly and crisply, results for this ten word vocabulary would be very good.

After testing the display as discussed above, it was evident that this scheme as it now stands is not sufficient for dependable computer recognition of speech. There is enough information available in these patterns to render them far from valueless, and to add support to the theory of interconnection of the zero-crossing rate and formant frequencies, but more information is needed for errorless identification of speech.

Better results would seem possible for this type comparison if an average of several voices were used as a dictionary. This would tend to minimize particular voice characteristics and accentuate the general sameness of the words being spoken. Time did not permit me to explore this possibility. It is recommended to anyone who wishes to pursue this work further.

If one is to hold the theory that the formants contain the information, and further, that the zero-crossing rate is a measure of the formant frequencies, then it is logical to say that sufficient information is available here for error-free speech identification, and the problem lies in the way this information is being handled. It was not proposed at the outset of this work that an Euclidean comparison of the patterns was the optimal way of performing recognition, and the results tend to say that it is far from satisfactory. Since the speech

signals are statistical in nature, it is reasonable to expect that any comparison system that does not allow for such a nature is not going to be satisfactory. It is proposed by this researcher that statistical methods be employed in future comparison systems.

The addition of another speech parameter to produce a four dimensional plot (formant one, formant two, time and some other parameter) is also worthy of consideration. While difficult to visualize, a four dimensional plot would be no problem to implement on even a computer as small as the one used for this project.

Satisfactory performance of a speech recognition scheme for even as small a vocabulary as the numbers zero through nine would have possible applications today. An example is the verification of credit card validity by a business. It has been my experience to observe that few businesses check their list of invalid credit cards, save for the first page or so, obviously taking the attitude that if the user has not run out as soon as they approach the list, then his card is probably good. A rapid telephone checking system could be realized by the businessman calling a preassigned number which would connect him with the computer listing of lost or stolen cards. By reading off the numbers he could have very rapid, current knowledge of the status of the card in question. Such a system could save substantial losses that are occurring presently.

In conclusion, it seems that the first and second formant analogs, as extracted from the speech sounds here, are a worthy measure of the intelligence being transferred, and more work in their processing is warranted. Using the computer to handle the pattern comparisons, and,

with the "error display" for visual monitoring of the computer's operation, more sophisticated comparison methods should improve on the results obtained thus far.

BIBLIOGRAPHY

1. Ainsworth, W., "Relative Intelligibility of Different Transforms of Clipped Speech," *Journal of the Acoustical Society of America*, Vol. 41, May 1967, pp. 1272-1276.
2. Barrett, N. A., "Extracting Analogue Signals from Noise Using a Digital Computer," M.S. Thesis, Naval Postgraduate School, Monterey, California, May 1966.
3. Bezdel, W. and Chandler, H. J., "Results of an Analysis and Recognition of Vowels by Computer using Zero-Crossing Data," *Proceedings of the Institute of Electrical Engineers (London)*, Vol. 112, November 1965.
- ✓4. Biddulph, R., "Short Term Autocorrelation Analysis and Correlationgrams of Spoken Digits," *Journal of the Acoustical Society of America*, Vol. 26, July 1954, pp. 539-541.
5. Chang, S., Pihl, G. E., and Essigmann, M. W., "Representations of Speech and Sounds and Some of Their Statistical Properties," *Proceedings of the Institute of Radio Engineers*, Vol. 39, February 1951, pp. 147-153.
- ✓6. Denes, P. and Mathews, M., "Spoken Digit Recognition Using Time-Frequency Pattern Matching," *Journal of the Acoustical Society of America*, Vol. 32, November 1960, pp. 1450-1455.
7. Dietrich, W., "Calculations of the Effects of Peak Clipping on Speech-Like Signals," M.S. Thesis, Naval Postgraduate School, Monterey, California, December 1966.
8. Fant, C., Acoustical Theory of Speech Production. The Hague: Mouton and Co., 1960.
9. Fairbanks, G., Voice and Articulation Drillbook. New York, Harper and Rowe Co., 1952.
10. Fletcher, H., Speech and Hearing in Communications. New York: D. Van Nostrand Company, 1953.
11. Gold, B. and Rader, C., "Systems for Compressing the Bandwidth of Speech," *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-15, September 1967, pp. 131-136.
- ✓12. Hollabaugh, J., "Methods for Phonemic Recognition in Speech Processing," M.S. Thesis, Naval Postgraduate School, Monterey, California, 1963.
13. Huddy, N. W., Jr., "An Investigation of Methods of Improving the Intelligibility of Audio Frequency Speech in Noise." M.S. Thesis, Naval Postgraduate School, Monterey, California, October 1966.

14. Hughes, G., "The Recognition of Speech by Machine," Technical Report No. 395, Research Laboratory of Electronics, MIT, Cambridge, Massachusetts, May 1961.
15. Kinsler, L. E. and Frey, A. R., Fundamentals of Acoustics. New York: John Wiley & Sons, Inc. 1950.
16. Licklider, J. and Pollack, I., "Effects of Differentiation, Integration, and Infinite Peak Clipping on the Intelligibility of Speech," Journal of the Acoustical Society of America, Vol. 20, January 1948, pp. 42-51.
17. Martin, T., Nelson, A., and Zadell, H., "Speech Recognition by Feature-Abstraction Techniques," Technical Documentary Report No. AL TDR 64-176, Radio Corporation of America, Camden, New Jersey, August 1964.
18. Olson, H. and Belar, H., "Phonetic Typewriter III," Journal of the Acoustical Society of America, Vol. 33, November 1961, pp. 1610-1615.
19. Olson, H. and Belar, H., "Printout System for the Automatic Recording of the Spectral Analysis of Spoken Syllables," Journal of the Acoustical Society of America, Vol. 34, February 1962, pp. 166-171.
20. Olson, H., Belar, H. and Rogers, E., "Speech Processing Techniques and Applications," IEEE Transactions on Audio and Electroacoustics, Vol. AU-15, September 1967, pp. 120-126.
21. Potter, R. K., Kopp, G. A., and Kopp, H. G., Visible Speech. New York, Dover Publications, 1966.
22. Pyron, B. and Williams, F., Jr., "Signal Processing by Infinite Clipping and Related Techniques," Final Report, Project A-727, U.S. Government Contract DA 49-092-AR0-21, Georgia Institute of Technology, Atlanta, Georgia, April 1964.
23. Sakai, T. and Doshita, S., "The Automatic Speech Recognition System for Conversational Sound," IEEE Transactions on Electronic Computers, Vol. EC-12, December 1963, pp. 835-846.
24. Sakai, T. and Inoue, S., "New Instruments and Methods for Speech Analysis," Journal of the Acoustical Society of America, Vol. 32, April 1960, pp. 441-450.
25. Sebestyen, G. S., Decision-Making Processes in Pattern Recognition, New York: Macmillan Co., 1962.
26. Sholtz, P. and Bakis, R., "Spoken Digit Recognition Using Vowel-Consonant Segmentation," Journal of the Acoustical Society of America, Vol. 34, January 1962, pp. 1-5.

- ✓ 27. Shoup, J., "Phoneme Selection for Studies in Automatic Speech Recognition," Journal of the Acoustical Society of America, Vol. 34, April 1962, pp. 397-403.
28. Teacher, C., Kellett, H. and Focht, L., "Experimental, Limited Vocabulary, Speech Recognizer," IEEE Transactions on Audio and Electroacoustics, Vol. AU-15, September 1967, pp. 127-130.
29. Wilde, J., "A Speech Analysis and Compression Scheme for Bandwidth Reduction," M.S. Thesis, Naval Postgraduate School, Monterey, California, June 1959.
30. Williams, D., "A Visual Display of Certain Speech Parameters," M.S. Thesis, Naval Postgraduate School, Monterey, California, 1967.

APPENDIX I

PRESENTATION AND DISCUSSION OF PYRON AND WILLIAMSON'S VECTOR DISPLAY EQUIPMENT

Figure 9 shows a block diagram of the circuitry used by Pyron and Williamson (Ref. 22) in producing their "vector display." This circuitry was reproduced in the course of this study (See Section 6).

The input voice sound is clipped, amplified, and clipped repeatedly in the infinite clipper until the waveshape is virtually rectangular. A Schmidt trigger completes the clipper action, yielding a wave whose only information is the zero-crossing points of the original waveform. This signal is then differentiated to give sharp pulses at the points of zero-crossing. These pulses, in turn, trigger the monostable multivibrator and a rectangular pulse occurs at the output corresponding to the zero-crossing. The time averaging produces a slowly varying analog proportional to the frequency of zero-crossing of the input. If the input is differentiated prior to processing, the analog will be proportional to the rate of maxima and minima that occurred in the original waveform.

The amplitude analog circuit half-wave rectifies the original waveform and forms a smoothed, slowly varying output proportional to the amplitude of the original speech signal.

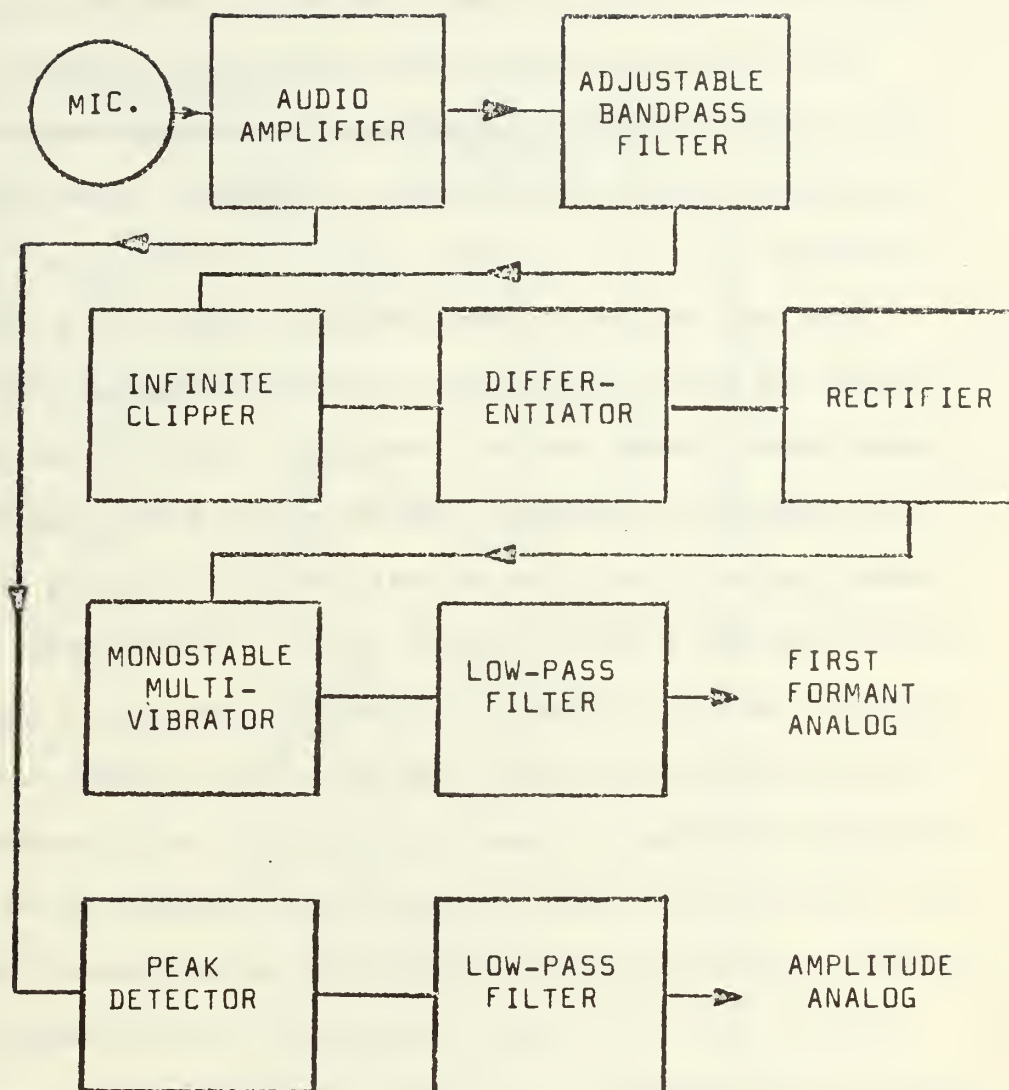


Figure 9. Block diagram of circuits used to produce the Vector Display.

APPENDIX II

CIRCUITS USED TO EXTRACT FORMANTS ONE AND TWO FROM THE SPEECH SIGNAL

The circuits described in this appendix were designed to extract from a speech wave analog signals which are proportional to the first and second formants. The operation of the second formant channel is as follows:

After the emitter follower input (see Figure 10) a differentiator provides six db per octave higher frequency preemphasis to bring the weaker second formant into the foreground. The differentiator feeds into a common base transistor used to provide a low impedance load, thereby improving the differentiating action. Following the next emitter follower there is a common emitter amplifier stage biased near cut-off to provide clipping on the positive side of the signal. The differentiator and diode that come next yield pulses at each positive going zero-crossing to trigger the mono-stable multivibrator (Figure 11). The monostable output pulses are then averaged by the low-pass filter to provide a voltage analog of the second formant frequency.

The first formant channel operates in a similar manner, with the same circuit, except for the omission of the six db per octave higher frequency preemphasis of the first differentiator.

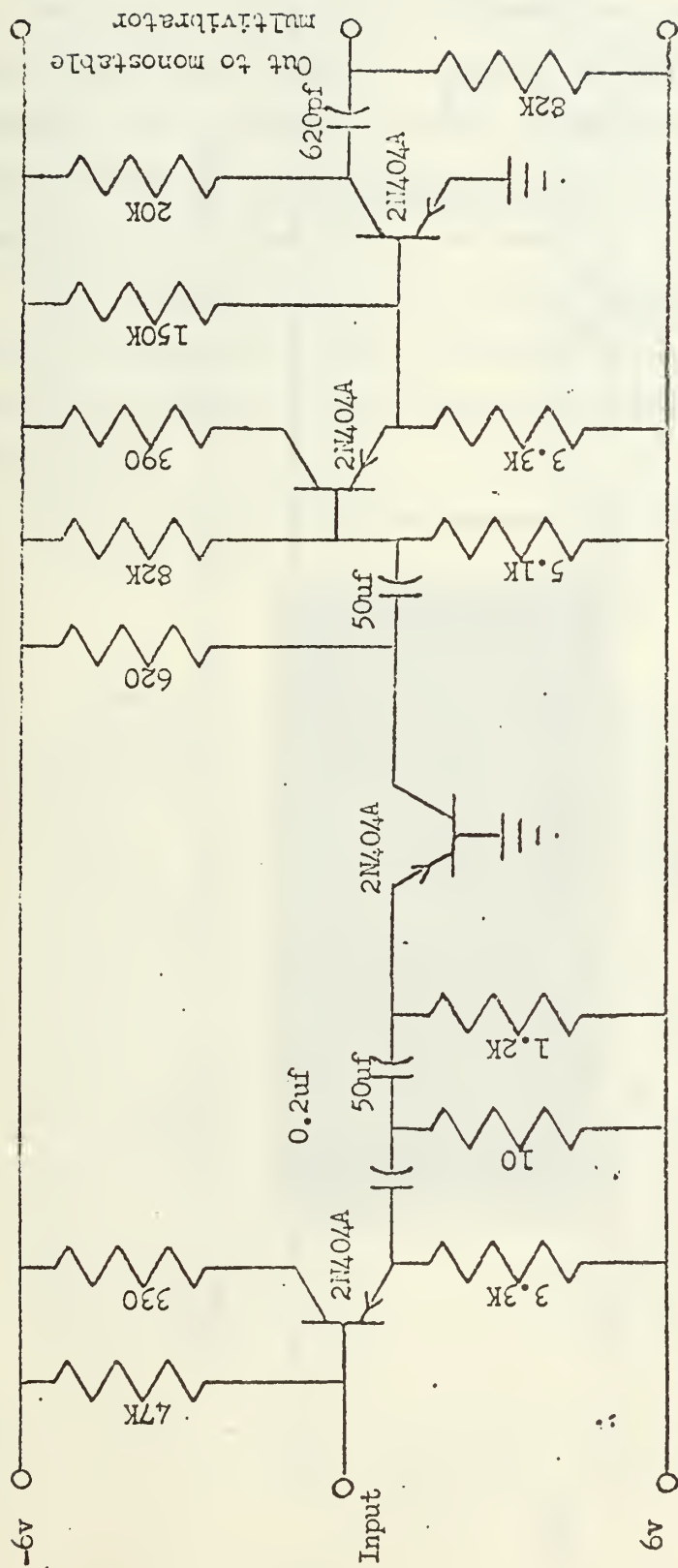


Figure 10. Circuit for generation of second formant frequency (continued on next page).

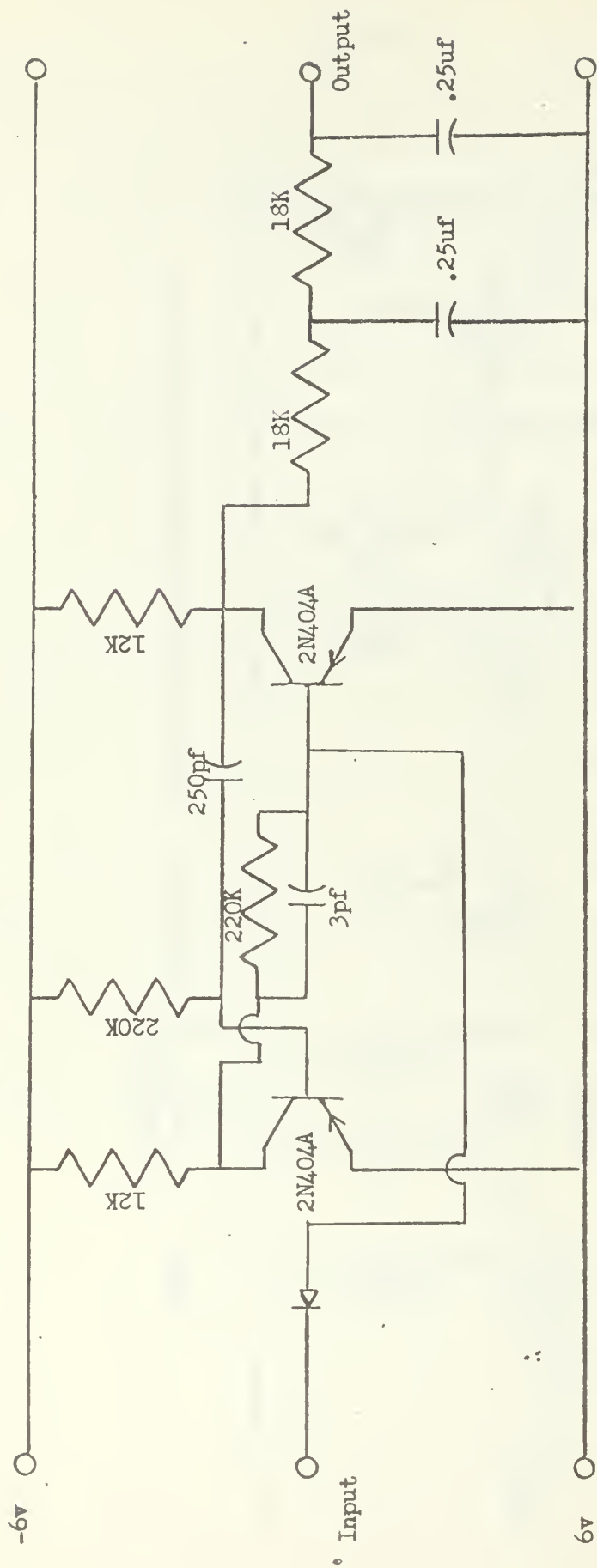


Figure 11. Monostable multivibrator and low-pass filter as used on both formant channels.

APPENDIX III

The patterns contained in this appendix were generated with the circuitry discussed in Section 7. They are typical for this type processing, and very subject to circuit parameters such as gain, bandwidth and degree of output averaging used. All shown here were generated under the same circuit conditions.

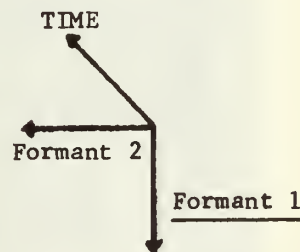
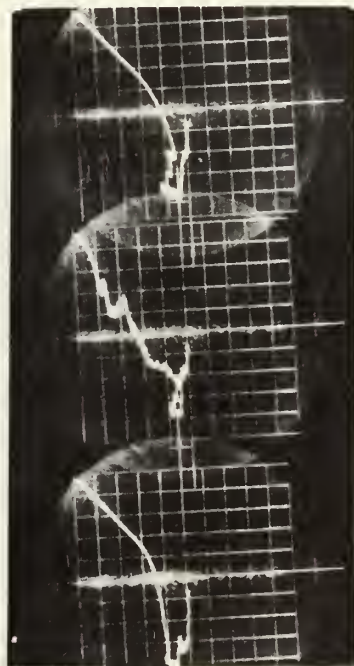
Each photograph contains three patterns of the same number spoken by three different male speakers. The order of speakers and the direction of components as shown below are the same for all the pictures in this section.

Number 1

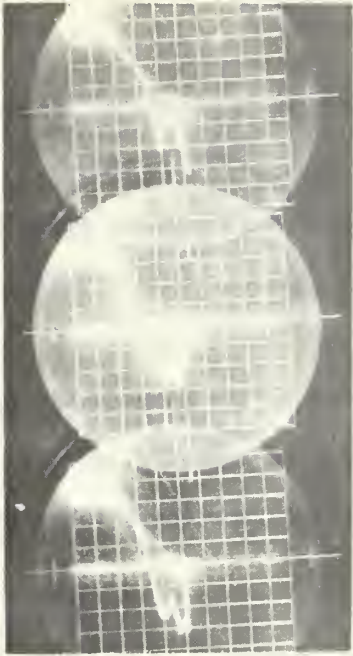
(RJL)

(TJK)

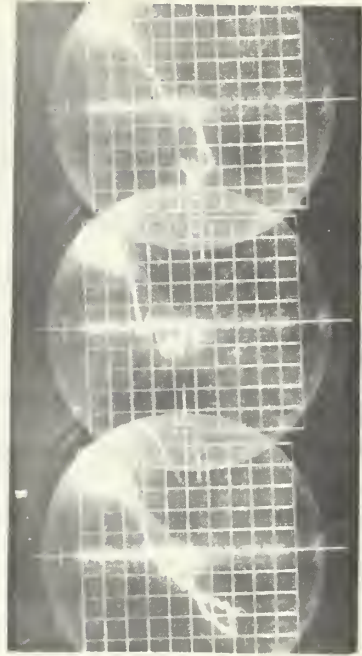
(HEK)



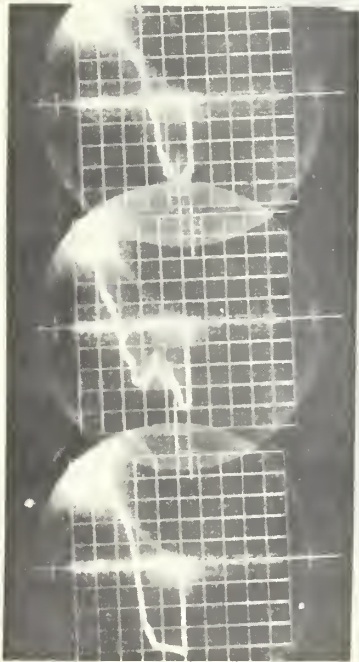
2



3



4

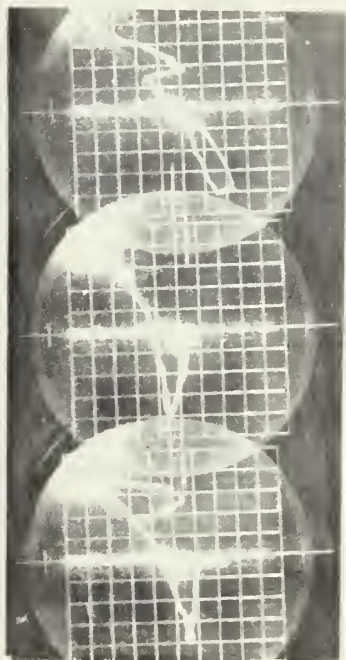


5

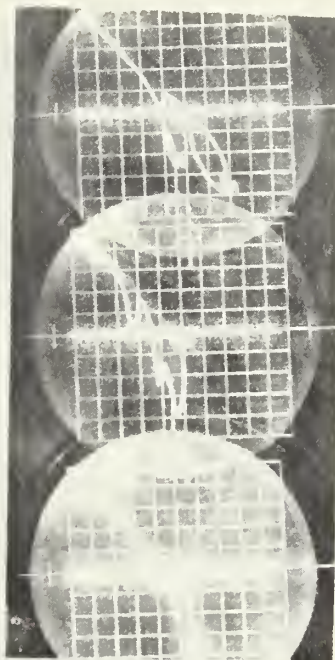


Figure 12. Formant 1 versus formant 2 versus time (diagonal) for numbers 2 through 5.

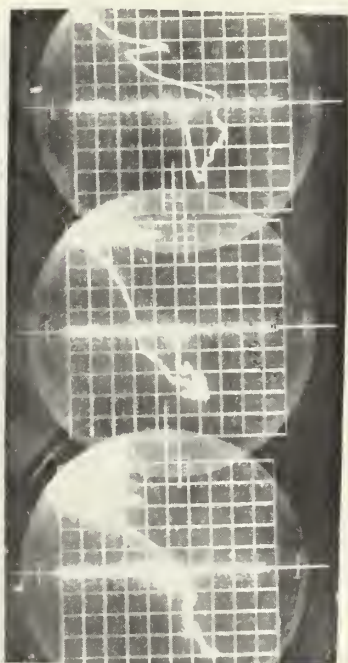
6.



7.



8.



9.

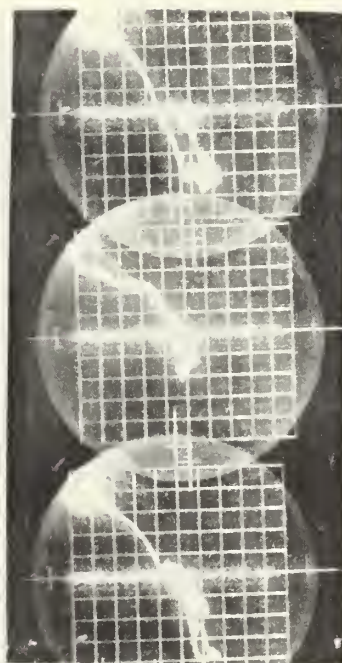
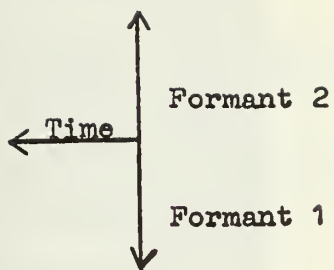


Figure 13. Formant 1 versus formant 2 versus time.
for numbers 6 through 9.

APPENDIX IV

The patterns shown in this appendix were generated by the circuits outlined in Appendix II and discussed in Section 8. All were made under the same circuit conditions and in the same environment. These are the type patterns that were used in attempting recognition of speech by computer in Sections 9 and 10.

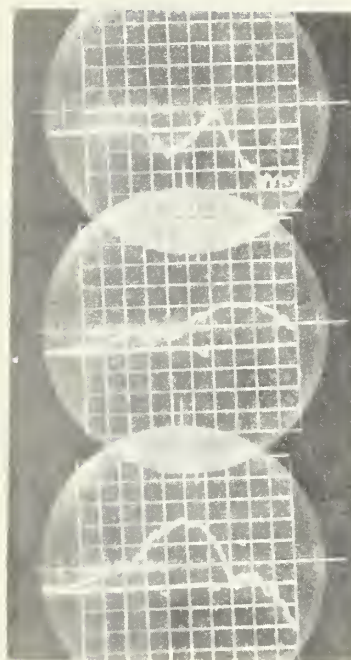


Numbers 1 to 9
by one male
speaker (TJK).

1

2

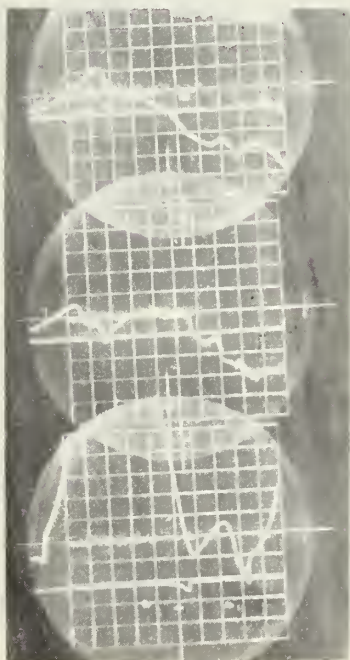
3



4

5

6



7

8

9

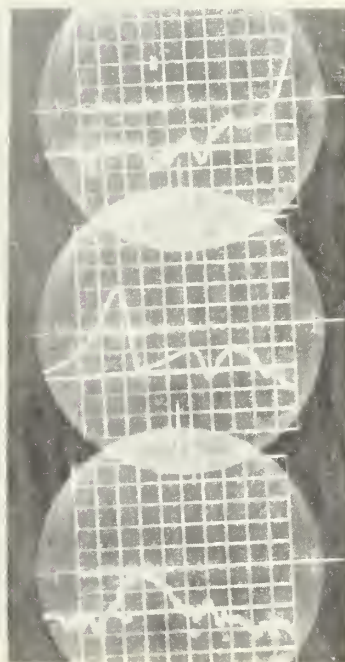
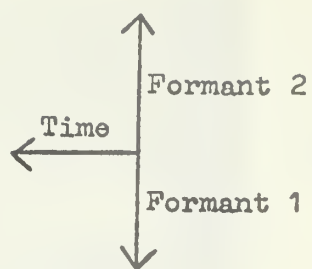


Figure 14. Formant 2 minus formant 1 versus time.



Numbers 1 to 9
by one male
speaker (RJL)

1

2

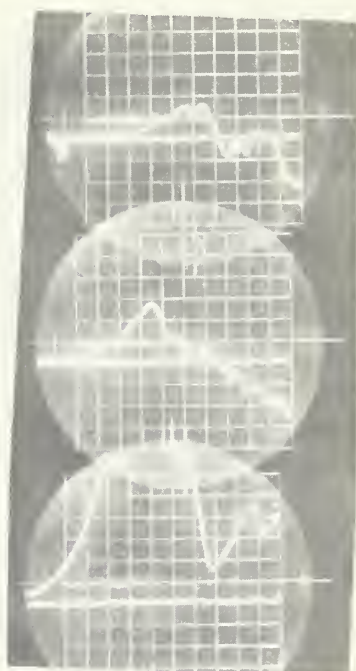
3



4

5

6



7

8

9

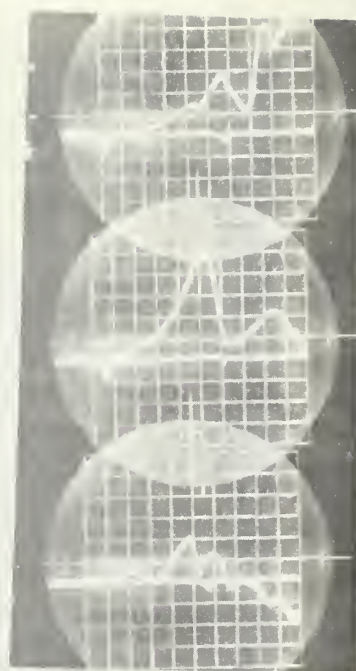
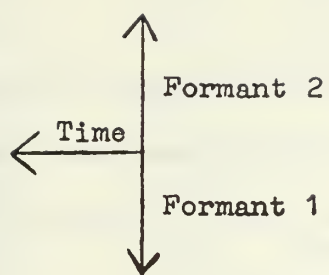


Figure 15. Formant 2 minus formant 1 versus time.



Numbers 1 to 9
by one male
speaker (HEK)

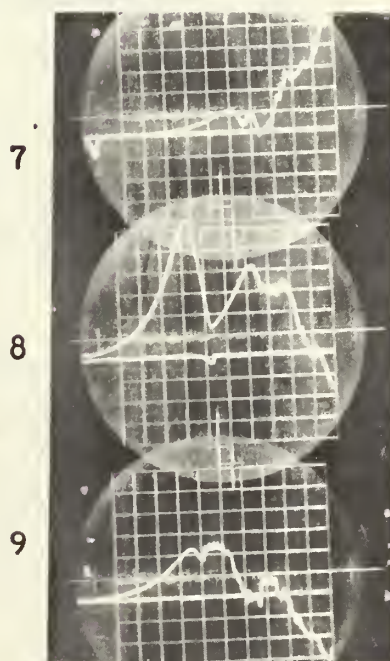
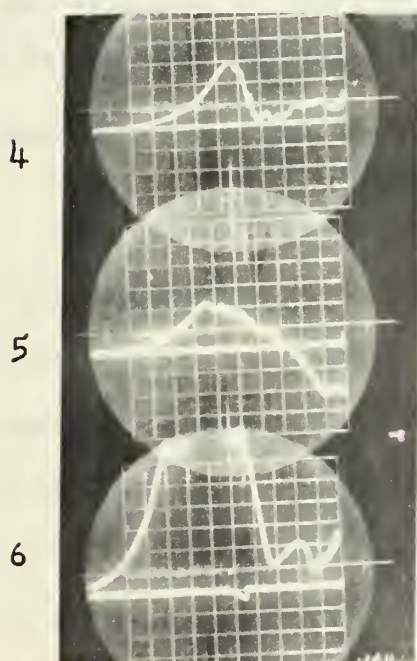
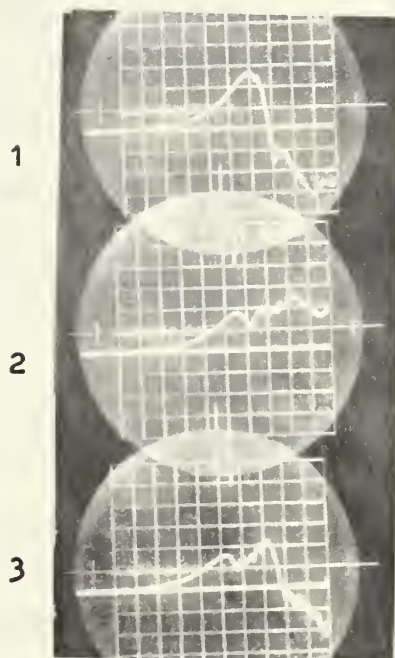


Figure 16. Formant 2 minus formant 1 versus time.

APPENDIX V

COMPUTER PROGRAMS

1. Program DICTIONARY.

This program is designed to take 600 samples from the inputs to channels 1 and 2 on the analog to digital converter, subtract 1 from 2 and store the difference on magnetic tape for future use.

Cell	Contents	Code	Explanations
0000	0101	PTA	Puts zero in A register
	4071	STD71	Puts zero in cell 71
	2200	LDC	Loads the following number in the register
	0600	0600	
	4070	STD70	Puts 600 in cell 70
	7500	EXF00	
	1401	1401	Call A/D channel 1
	7600	INA	Input
0010	4063	STD63	Store sample in cell 63
	2463	LCD63	Load complement of sample in the a register
	3642	SBF42	Subtract threshold value stored 42 cells ahead
	6704	NJB04	If threshold not exceeded go back and sample again
	6206	PJF06	
	7500	EXF00	
	1401	1401	
	7600	INA	
0020	4063	STD63	
	2463	LCD63	
	4076	STD76	
	7500	EXF00	
	1402	1402	Call A/D channel 2
	7600	INA	
	4063	STD63	
	2463	LCD63	

Cell	Contents	Code	Explanation
0030	3476	SBD76	Channel 2 minus channel 1
	4151	STI51	Store difference in address contained in cell 51
	5617	AOF17	
	2620	LCF20	
	0601	ADN01	
	6501	NZB01	Time delay for proper sample spacing
	5471	AOD71	
	3470	SBD70	
0040	6717	NJB17	Go back and take next samples
	7500	EXF00	
	2111	2111	Call tape unit #1
	7304	Out	
	3131	3131	Last word address plus 1 of data for tape storage
	2200	LDC	
	2000	2000	First address of data for tape
	4202	STF02	Initialize for next run
0050	7700	HLT	Halt
	2000	2000	
	2000	2000	
	0071	0071	Time delay constant
	0122	0122	0122 = 0.2v threshold

2. Program COMPARE.

This program provides for a comparison of an unknown word with ten previously stored words on tape, and an error versus time plot on the DD-65 display unit for each of the ten comparisons. At the completion of the ten comparisons the closest comparing word location on tape is stored in cell 51.

Cell	Contents	Code
0100	2200	LDC
	2000	2000
	4243	STF43

Cell	Contents	Code
	2200	LDC
	4000	4000
	4202	STF02
	0400	LDN00
	4077	STD77
0110	5701	AOB01
	0277	LPN77
	3647	SBF47
	6705	NJB05
	0512	LCN12
	4067	STD67
	2200	LDC
	1130	1130
0120	4070	STD70
	7500	EXF00
	1401	1401
	7600	INA
	4063	STD63
	2463	LCD63
	3600	SBC
	0122	0122
0130	6705	NJB05
	6205	PJF05
	7500	EXF00
	1401	1401
	7600	INA
	4063	STD63
	2463	LCD63
	4076	STD76
0140	7500	EXF00
	1402	1402
	7600	INA
	4063	STD63

Cell	Contents	Code
	2200	LDF00
	2000	2000
	4077	STD77
	2463	LCD63
0150	3476	SBD76
	4177	STI77
	5705	AOB05
	2611	LCF11
	0601	ADN01
	6701	NJB01
	0407	LDN07
	6207	PJF07
0160	0000	0000
	0077	0077
	7000	7000
	0700	0700
	0071	0071
	0122	0122
	5471	AOD71
	3470	SBD70
0170	6736	NJB36
	2200	LDC
	5001	5001
	4236	STF36
	2200	LDC
	5002	5002
	4237	STF37
	2200	LDC
0200	5003	5003
	4241	STF41
	2200	LDC
	5004	5004
	4241	STF41
	0400	LDN00

Cell	Contents	Code
	4075	STD75
	7500	EXF00
0210	2131	2131
	7203	INP03
	4265	4265
	6102	NZF02
	3134	3134
	2100	LDM
	2000	2000
	3500	SBM
0220	3134	3134
	6205	PJF05
	2100	LDM
	3134	3134
	3500	SBM
	2000	2000
	7101	JF101
	6174	NZF74
0230	0110	LS03
	5051	RAD51
	2055	LDD55
	1350	LPB50
	0110	LS03
	5052	RAD52
	2055	LDD55
	0270	LPN70
0240	0111	LS06
	0110	LS03
	5053	RAD53
	2055	LDD55
	0207	LPN07
	5054	RAD54
	5730	AOB30
	5727	AOB27

Cell	Contents	Code
0250	5723	A0B23
	5726	A0B26
	5475	A0D75
	3600	SBC
	1000	1000
	6740	NJB40
	2200	LDC
	2000	2000
0260	4342	STB42
	4334	STB34
	2200	LDC
	3134	3134
	4344	STB44
	4342	STB42
	0404	LDN04
	5336	RAB36
0270	0404	LDN04
	5334	RAB34
	0404	LDN04
	5331	RAB31
	0404	LDN04
	5330	RAB30
	2067	LDD67
	7101	JFI01
0300	7700	HLT
	5614	A0F14
	0401	LDN01
	5000	RAD00
	2001	LDD01
	6105	NZF05
	2000	LDD00
	4065	STD65
0310	5701	A) B01
	5457	AOD57
	0404	LDN04

Cell	Contents	Code
	5307	RAB07
	0411	LDN11
	0701	SBN01
	6615	PJB15
	2200	LDC
0320	2001	2001
	4315	STB15
	2200	LDC
	4065	4065
	4315	STB15
	2200	LDC
	0701	0701
	4312	STB12
0330	2065	LDD65
	6142	NZF42
	0401	LDN01
	5051	RAD51
	5622	AOF22
	0401	LDN01
	5053	RAD53
	2001	LDD01
0340	3405	SBD05
	6312	NJF12
	2302	LDB02
	0277	LPN77
	3200	ADC
	2000	2000
	4307	STB07
	2053	LDD53
0350	5051	RAD51
	0400	LDN00
	4053	STD53
	0404	LDN04
	5314	RAB14

Cell	Contents	Code
	0411	LDN11
	0701	SBN01
	6623	PJB23
0360	2200	LDC
	0701	0701
	4304	STB04
	2200	LDC
	3405	3405
	4325	STB25
	2200	LDC
	2001	2001
0370	4331	STB31
	0401	LDN01
	6264	PJF64
	2066	LDD66
	6060	ZJF60
	5615	AOF15
	2001	LDD01
	6110	NZF10
0400	2302	LDB02
	0601	ADN01
	0277	LPN77
	4030	STD30
	2130	LDI30
	4072	STD72
	5701	A0B01
	0404	LDN04
0410	5312	RAB12
	0411	LDN11
	0701	SBN01
	6616	PJB16
	2200	LDC
	4072	4072
	4311	STB11

Cell	Contents	Code
	2200	LDC
0420	2001	2001
	4323	STB23
	2200	LDC
	0701	0701
	4312	STB12
	2072	LDD72
	3473	SBD73
	6212	PJF12
0430	2065	LDD65
	4051	STD51
	0402	LDN02
	3457	SBD57
	6022	ZJF22
	2072	LDD72
	3474	SBD74
	6317	NJF17
0440	6211	PJF11
	2066	LDD66
	4051	STD51
	0402	LDN02
	3457	SBD57
	6011	ZJF11
	2074	LDD74
	3473	SBD73
0450	6206	PJF06
	2067	LDD67
	4051	STD51
	6203	PJF03
	2065	LDD65
	4051	STD51
	7500	EXF00
	1121	1121

Cell	Contents	Code
0460	2051	LDD51
	7101	JFI01
	7600	INA
	2066	LDD66
	4051	STD51
	0402	LDN02
	3457	SBD57
	6011	ZJF11
0470	2074	LDD74
	3473	SBD73
	6206	PJF06
	2067	LDD67
	4051	STD51
	6203	PJF03
	2065	LDD65
	4051	STD51
0500	7500	EXC
	1121	LPI21
	2051	LDD51
	7700	HLT

INITIAL DISTRIBUTION LIST

	No. of Copies
1. Defense Documentation Center Cameron Station Alexandria, Virginia 72314	20
2. Library Naval Postgraduate School Monterey, California 93940	2
3. Dr. G. D. Ewing Department of Electrical Engineering Naval Postgraduate School Monterey, California 93940	2
4. Professor D. B. Hoisington Department of Electrical Engineering Naval Postgraduate School Monterey, California 93940	1
5. Professor H. A. Titus Department of Electrical Engineering Naval Postgraduate School Monterey, California 93940	1
6. LT J. F. Taylor, USN 384D Bergin Drive Monterey, California 93940	1

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Naval Postgraduate School Monterey, California 93940		2a. REPORT SECURITY CLASSIFICATION Unclassified	
		2b. GROUP	
3. REPORT TITLE Computer Recognition of Speech Utilizing Zero-Crossing Information			
4. DESCRIPTIVE NOTES (Type of report and, inclusive dates) Thesis (Degree of Electrical Engineer)			
5. AUTHOR(S) (First name, middle initial, last name) John F. TAYLOR			
6. REPORT DATE June 1968		7a. TOTAL NO. OF PAGES 65	7b. NO. OF REFS 30
8a. CONTRACT OR GRANT NO.		9a. ORIGINATOR'S REPORT NUMBER(S)	
b. PROJECT NO.			
c.		9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	
d. <i>Unlimited Distribution</i>			
10. DISTRIBUTION STATEMENT This document is subject to special export controls and each transmittal to foreign nations may be made only with prior approval of the Naval Postgraduate School.			
11. SUPPLEMENTARY NOTES N/A		12. SPONSORING MILITARY ACTIVITY Naval Postgraduate School Monterey, California	
13. ABSTRACT <p>The nature of speech sounds is studied with particular emphasis on the information bearing elements of speech. The association of the amplitude clipped-speech zero-crossing rate, formant frequencies and information content of a speech signal is presented and capitalized upon to produce readily extractable first and second formants from the speech wave.</p> <p>Various methods of processing the formants to generate unique patterns for particular sounds are attempted, with a time plot of the arithmetic difference of the two formants being explored in detail. The object being to obtain machine recognition of speech.</p> <p>Control Data Corporation 160 computer machine language programs are prepared to realize an Euclidean comparison of spoken numbers zero to nine against a previously stored "dictionary." Testing showed this type processing satisfactory for some voices, but not readily extendible to many voices with the same dictionary." Methods of overcoming this shortcoming are suggested.</p>			

KEY WORDS

LINK A

LINK B

LINK C

ROLE

WT

ROLE

WT

ROLE

WT

Clipped speech

~~NO FORN~~

Thesis ~~NO FORN~~ 107780
T22143 Taylor
c.1 Computer recognition
of speech utilizing
zero-crossing informa-
tion.
12 CT 69 S10069
14 SEP 70 18640
12 SEP 88 55269

Thesis 107780
T22143 Taylor
c.1 Computer recognition
of speech utilizing
zero-crossing informa-
tion.

thesT22143

Computer recognition of speech utilizing



3 2768 001 01071 3

DUDLEY KNOX LIBRARY